Pattern recognition techniques and their applications for automatic classification of artificial partial discharge sources

Pattern Recognition Techniques and Their Applications for Automatic Classification of Artificial Partial Discharge Sources

Hui Ma, Jeffery C. Chan, Tapan K. Saha and **Chandima Ekanayake**
The University of Queensland
Brisbane, Australia

ABSTRACT

Partial discharge (PD) source classification aims to identify the types of defects causing discharges in high voltage (HV) equipment. This paper presents a comprehensive study of applying pattern recognition techniques to automatic PD source classification. Three challenging issues are investigated in this paper. The first issue is the feature extraction for obtaining representative attributes from the original PD measurement data. Several approaches including stochastic neighbour embedding (SNE), principle component analysis (PCA), kernel principle component analysis (KPCA), discrete wavelet transform (DWT), and conventional statistic operators are adopted for feature extraction. The second issue is the pattern recognition algorithms for identifying various types of PD sources. A novel fuzzy support vector machine (FSVM) and a variety of artificial neural networks (ANNs) are applied in the paper. The third issue is the identification of multiple PD sources, which may occur in HV equipment simultaneously. Two approaches are proposed to address this issue. To evaluate the performance of various algorithms in this paper, extensive laboratory experiments on a number of artificial PD models are conducted. The classification results reveal that FSVM significantly outperforms a number of ANN algorithms. The practical PD sources classification for HV equipment is a considerable complicated problem. Therefore, this paper also discusses some issues of meaningful application of the above proposed pattern recognition techniques for practical PD sources classification of HV equipment.

Index Terms — **Partial discharge (PD), PD source classification, artificial neural network (ANN), fuzzy support vector machine (FSVM), and pattern recognition.**

1 INTRODUCTION

PARTIAL discharge (PD) is a localized breakdown in the insulation system of high voltage (HV) equipment. Over the last few decades, PD phenomena has attracted extensive investigations related to its physical and chemical mechanisms, detection and acquisition techniques, de-nosing and filtering methods, and discharge source classification [1-23]. Among these investigations, automatic PD source classification has significant benefits for condition assessment of in-service HV equipment since it can reveal the types of defects (PD sources) that cause discharge [1-7, 9-13, 16, 18-23].

It is well-known that different types of PD sources can generate different discharge patterns [1-7, 10, 13]. The discharge pattern is represented by a set of characteristic attributes (features), which can be extracted from PD measurement data. Two sets of features are commonly adopted in PD source classification. They are the phase resolved discharge pulses' magnitude and repetition rate distribution, and the time resolved discharge pulses shape and magnitude [1-7, 10, 13]. On the basis of these features, it is possible to develop computer algorithms for automatically classifying the types of PD sources to assist condition assessment of HV equipment. With the advancement of digital electronics and signal processing techniques, considerable efforts have been made to apply various artificial intelligence (AI) techniques such as artificial neural networks (ANNs), genetic algorithms, knowledge-based system, fractal models, wavelet transformation, and support vector machines (SVMs) for automatic PD source classification [1-7, 9-13, 16, 18-23].

However, there are still considerable challenges remaining for successfully applying AI techniques to PD source classification. Three key challenges are: (1) extracting suitable features from raw data obtained from PD measurement; (2) applying suitable pattern recognition algorithms to PD source classification; and (3) recognizing multiple PD sources, which may occur in HV equipment simultaneously.

This paper investigates pattern recognition techniques and their applications for automatic PD source classification. To

Manuscript received on 14 October 2012, in final form 23 January 2013.

address the essential requirements of PD source classification, a pattern recognition application framework is presented in the paper. To obtain proper features from PD measurement data, this paper adopts a number of approaches including statistic operators, principle component analysis (PCA), kernel principle component analysis (KPCA), stochastic neighbour embedding (SNE), and discrete wavelet transform (DWT). In order to investigate the applicability of different pattern recognition algorithms for PD source classification, this paper implements various algorithms including a novel fuzzy support vector machine (FSVM) and a number of artificial neural networks (ANN) such as radial basis network (RBF), multi-layer perceptron (MLP), Bayesian classifier, *k* nearest neighbour (*K*NN), and two-layer (input-output) network. Moreover, two approaches have been proposed for identifying multiple PD sources, which may occur in HV equipment simultaneously. Extensive laboratorial experiments are conducted by using artificial PD models. And a comprehensive performance evaluation and comparison on various algorithms developed in this paper are presented. This paper also discusses a number of practical issues for meaningful application of the proposed pattern recognition techniques to PD sources classification of HV equipment.

2 PATTERN RECOGNITION FRAMEWORK FOR PD SOURCE CLASSIFICATION

As shown in Figure 1, PD source classification consists of three major stages: (1) data acquisition and pre-processing, which captures, digitizes, and purifies PD signal; (2) feature extraction, which extracts the representative attributes from raw PD data; and (3) pattern recognition, which identifies the types of PD sources. This paper focuses on the above stages (2) and (3). In feature extraction stage, the measurement data is converted to a phase resolved dataset, which consists of PD pulse magnitude and pulse number distribution with respect to the phase of power cycle (Figure 1a); and then the representative attributes are extracted from this dataset (Figure 1b). In pattern recognition stage, fuzzy support vector machine (FSVM) or artificial neural networks (ANNs) are applied to recognize various PD sources (Figures 1c and 1d).

The mathematic formulation of PD source classification is as follows. After data acquisition and pre-processing, the resultant PD data is a N x D dataset $\mathbf{X} = [\,\mathbf{x}_1, \ldots, \mathbf{x}_N\,]$, where N is the number of data points in the dataset, and each data point is a D dimensional vector, i.e. $\mathbf{x}_i = [x_i^1, \ldots, x_i^D]$. The feature extraction stage transforms the dataset $\mathbf{X}$ to a N x d dataset $\mathbf{Y} = [\,\mathbf{y}_1, \ldots, \mathbf{y}_N\,]$, $\mathbf{y}_i = [y_i^1, \ldots, y_i^d]$, where $d < D$. Though substantial features are removed from the dataset $\mathbf{X}$ during the feature extraction stage, the remaining features in the dataset $\mathbf{Y}$ can still provide essential characteristics for discriminating different types of PD sources. This will be discussed in the next section. Each data point in dataset $\mathbf{X}$ or $\mathbf{Y}$ belongs to one of the T independent classes, $C_k \in \{1, \ldots, T\}$, where each class corresponds to one type of PD sources. The task for pattern recognition algorithm is to learn the underlying mathematical relationship between data points and their corresponding types of PD sources, i.e. to explore the function of $\mathbf{f}: \mathbf{y}_j \rightarrow C_k,\ j = 1, 2, \ldots N,\ C_k \in \{1, 2, \ldots T\}$. Such knowledge can then be used to classify a new data point to its belonged type of PD source.

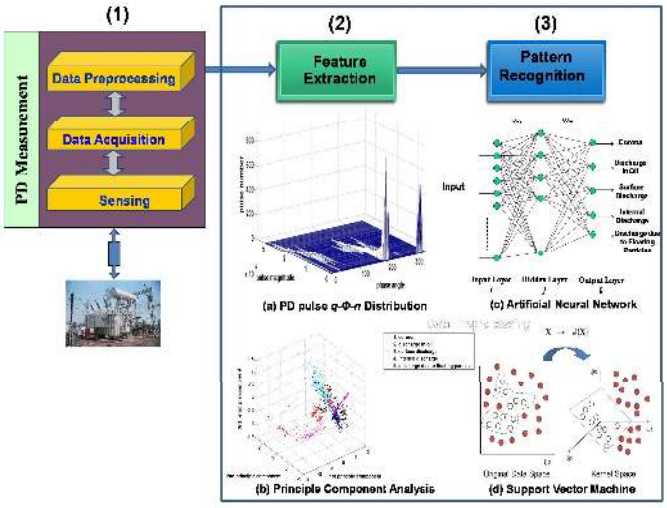


Figure 1. Pattern recognition techniques for PD source classification (a) PD *q*-*Φ*-*n* distribution; (b) Principle component analysis (PCA) on PD dataset; (c) Three-layer artificial neural network (ANN); (d) Illustrative concept of support vector machine (SVM).

3 FEATURE EXTRACTION FOR PD SOURCE CLASSIFICATION

This paper applies a number of feature extraction approaches including statistic operators, principle component analysis (PCA), kernel principle component analysis (KPCA), stochastic neighbour embedding (SNE), and discrete wavelet transform (DWT) for PD source classification.

The statistic operators, PCA, KPCA, and SNE extract features from the phase resolved PD data, in which the power frequency cycle of the applied voltage is divided into a number of phase windows and PD pulses are quantified by their average magnitude, maximum magnitude, and number count in each phase window [4]. The phase resolved dataset has a considerable high dimensionality. For a PD dataset of 200 phase windows (1.8˚ per window of a power frequency cycle), there will be $200 \times 3 = 600$ features per data point. The statistic operators, PCA, KPCA, and SNE can deal with such high dimensionality and extract representative features for PD source classification. Instead of using the phase resolved PD data, DWT utilizes original PD signals for feature extraction. The original PD signals are represented by a set of wavelet approximation coefficients at different scales. The first four moments of the probability distribution of these approximation coefficients are adopted to form the feature set.

This section provides a brief review on statistic operators, PCA, KPCA, SNE, and DWT. Due to the space limitation; the mathematical derivations are kept short.

3.1 STATISTIC OPERATORS

In this approach, a number of statistic operators are calculated for the phase resolved PD pulse distributions including maximum pulse magnitude distribution, average

pulse magnitude distribution, and pulse number distribution [4]. The feature set consists of total 24 features as follows:

(1) Skewness and Kutrosis for both positive and negative cycles of the above phase resolved distributions.
(2) Number of peaks for both positive and negative cycles of the above phase resolved distributions.
(3) Discharge asymmetry between positive and negative cycles of the above phase resolved distributions.
(4) Cross-correlation factor of the above phase resolved distributions.

3.2 PRINCPLE COMPOMENT ANAYSIS (PCA)

PCA embeds the original data into a lower dimensional linear subspace, which preserves as much of the variance in the data as possible [29]. In order to seek this linear subspace, PCA solves the following eigen problem:

$$cov(\mathbf{X})\mathbf{M} = \boldsymbol{\lambda}\mathbf{M} \tag{1}$$

where $cov(\mathbf{X})$ is the covariance matrix of the dataset $\mathbf{X}$, $\mathbf{M}$ is a linear mapping formed by the d principle eigenvectors of the covariance matrix, and $\boldsymbol{\lambda}$ are the d principal eigenvalues. The low-dimensional data representations $\mathbf{y}_i$ of the data points $\mathbf{x}_i$ are computed through a linear mapping $\mathbf{Y} = \mathbf{XM}$. The elements of $\mathbf{Y}$, e.g. $\mathbf{y}_i$ will form the features set.

3.3 KERNEL PCA (KPCA)

KPCA is one of the nonlinear feature extraction methods [29]. It firstly uses a kernel function to transform the original data into a feature space. The elements of this kernel is in the form of

$$K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \tag{2}$$

where κ is a kernel function, $\mathbf{x}_i$ and $\mathbf{x}_j$ are different data points in dataset $\mathbf{X}$. One commonly used kernel is Gaussian kernel (refer to Section 4). KPCA computes the eigenvectors of the covariance matrix $\mathbf{a}_i$ in the above feature space.

To obtain the low-dimensional data representation $\mathbf{Y}$, the data points in the original dataset $\mathbf{X}$ are projected onto the eigenvectors of the covariance matrix $\mathbf{a}_i$ as

$$\mathbf{y}_i = \left\{\sum_{j=1}^{N} a_1^{(j)} \kappa(\mathbf{x}_i, \mathbf{x}_j), \ldots, \sum_{j=1}^{N} a_d^{(j)} \kappa(\mathbf{x}_i, \mathbf{x}_j)\right\} \tag{3}$$

As in PCA, the elements $\mathbf{y}_i$ will form the features set.

3.4 STOCHASTIC NEIGHBOR EMBEDDING (SNE)

SNE attempts to retain the pairwise distances between the data points in the low-dimensional space [27]. In SNE, the probability p_{ij} that data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are generated by the same Gaussian is computed for all possible pairs of data points in the original dataset $\mathbf{X}$. The probabilities q_{ij} of the corresponding data points in low dimensional space, i.e. $\mathbf{y}_i$ and $\mathbf{y}_j$ generated by the same Gaussian are also computed. These two probability matrices are denoted as $\mathbf{P}$ and $\mathbf{Q}$, respectively. SNE minimizes the difference between the probability distributions $\mathbf{P}$ and $\mathbf{Q}$ through minimizing the sum of Kullback-Leibler divergences

$$\varphi = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{4}$$

The minimization of φ can be performed using a gradient descent method. As in PCA and KPCA, the elements $\mathbf{y}_i$ will form the features set.

In PCA, KPCA and SNE, the dimension of dataset $\mathbf{Y}$, d is called intrinsic dimensionality. It is the minimum number of features that is necessary to account for all information in the data. The intrinsic dimensionality of a dataset can be determined by using maximum likelihood estimation [24].

3.5 DISCRETE WAVELET TRANSFORM (DWT)

The advantages of using discrete wavelet transform (DWT) as a feature extraction tool is in that it can integrate PD signal de-nosing and feature extraction in a single step.

Wavelet is a small wave-type signal, which satisfies $\int_{-\infty}^{\infty} \psi(t)dt = 0$ and $\int_{-\infty}^{\infty} [\psi(t)]^2 dt < \infty$. A signal can be transformed into wavelet coefficients as [31]

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right) \tag{5}$$

where a is a scale factor and b is a translation factor. The scale factor is for compressing and stretching the mother wavelet while the translation factor is for shifting the mother wavelet along the time axis. The discrete wavelet transform (DWT) is defined as [31]

$$W_{j,k} = \sum_{n \in z} \varphi(n) 2^{(-j/2)} \psi(2^{-j} n - k) \tag{6}$$

where $\varphi(n)$ is the discrete function of a signal, j and k are integers. In the implementation of DWT, the original signal undergoes a series of low pass and high pass filters and is decomposed into a number of approximation and detail coefficients until reaching a predefined decomposition level.

To construct the features for PD source classification, the original PD signals (discharge pulses) are decomposed into nine levels. Consequently, each discharge pulse is represented by nine detail coefficients [13, 21]. Apparently, this introduces considerable high dimensionality as several hundred discharge pulses may generate in one power cycle. Thus, in this paper the first four moment statistics (i.e. mean, variance, skewness, and kurtosis) are computed for the probability distribution formed by the nine coefficients of PD pulses. Total 36 features ($4 \times 9 = 36$) per data point are constructed using DWT.

4 PATTERN RECOGNITION ALGORITHMS FOR PD SOURCES CLASSIFICATION

The pattern recognition algorithm generally involves two steps: training (learning) and testing (classification). In the training step, the task is to model the mathematic function between data points and their corresponding types of PD sources. In the testing step, a new input data point (not included in the training data) is classified into one type of PD sources [2, 3, 25]. This section investigates a novel fuzzy support vector machine (FSVM) and a variety of artificial neural networks (ANNs) for PD source classification. The mathematical formulations of these algorithms will be briefly introduced in this section. For the detailed derivations of these algorithms, readers may refer to [24, 26, 29].

4.1 *K*-NEAREST NEIGHBOUR (*K*NN)

In *K*NN, the distances between a new data point **y*** to each data point in the training dataset are computed. Let S_k denotes a subset consisting of the K closest data points and K_l denotes the frequency of the l-th class in S_k. Then **y*** is assigned to the class t, which has the largest frequency in S_k:

$$K_t \geq K_l \quad \text{for } l = 1, \dots, T \tag{7}$$

4.2 TWO-LAYER (INPUT AND OUTPUT) NETWORK

In two-layer network (input and output layer), a linear combination of input data points is firstly formed as [24]

$$a_i^k = \sum_{j=1}^{d} w_i^{kj} y_i^j + b_i^j \quad i = 1, \dots, N;\ k = 1, \dots, T \tag{8}$$

where y_i^j is the j-th element of the i-th data point $\mathbf{y}_i$, w_i^{kj} is the weight regarding the k-th output to the j-th element of $\mathbf{y}_i$, b_i^j is the bias term, N is the total number of data points, d is the data dimension (feature size), and T is the number of types of PD sources. Then a softmax activation function is applied to the above variables a_i^k to give output [24] as

$$z_i^k = \frac{exp(a_i^k)}{\sum_{k'=1}^{T} exp(a_i^{k'})} \qquad k = 1, \dots, T \tag{9}$$

An iterative reweighted least squares (IRLS) algorithm is used for training this two-layer network [24].

4.3 MULTI-LAYER PERCEPTRON (MLP)

MLP is a three layers (input layer, hidden layer, and output layer) network and consists of two sets of adaptive weights [24]. The network structure of MLP is shown in Figure 1c. A linear combination of data points in the input layer is formed to give a set of variables associated with the hidden layer

$$\alpha_i^m = \sum_{j=1}^{d} w_i^{mj} y_i^j + b_i^m \quad i = 1, \dots, N;\ m = 1, \dots, M \tag{10}$$

where y_i^j is the j-th element of the i-th data point $\mathbf{y}_i$, w_i^{mj} is the weight element regarding the m-th hidden node to the j-th element of data point $\mathbf{y}_i$, b_i^m is the bias term, and M is the number of hidden nodes. Then a tanh activation function is applied to the above variables α_i^m as

$$\gamma_i^m = tanh(\alpha_i^m) \qquad m = 1, \dots, M \tag{11}$$

The variables γ_i^m are then transformed by the second sets of weights regarding the k-th output to the m-th hidden node as

$$a_i^k = \sum_{m=1}^{M} w_i^{km} \gamma_i^m + b_i^k \quad i = 1, \dots, N;\ k = 1, \dots, T \tag{12}$$

Finally these values go through a soft-max activation function to give output z_i^k. The MLP is trained using the conventional back-propagation techniques [24].

4.4 RADIAL BASIS FUNCTION (RBF) NETWORK

The RBF network is different from MLP in that the activation function of the hidden nodes is based on a non-linear function, which is decided by the distance between the input vector and a weight vector. RBF network aims to derive an input-output mapping using M radial basis functions [29]

$$z(\mathbf{y}) = \sum_{m=1}^{M} w^{km} \phi_m(\mathbf{y}) \qquad k = 1, \dots, T \tag{13}$$

where ϕ_m are the basis functions and w^{km} are the output layer weights. The posterior probability of class membership is:

$$p(C_k|\mathbf{y}) = \sum_{m=1}^{M} w^{km} \phi_m(\mathbf{y}) \tag{14}$$

where $C_k \in \{1, \dots, T\}$, T is the number of classes, and the basis function ϕ_m are given by

$$\phi_m(\mathbf{y}) = \frac{p(\mathbf{y}|m)P(m)}{\sum_{m'=1}^{M} p(\mathbf{y}|m')P(m')} = P(m|\mathbf{y}) \tag{15}$$

and the weights are given by

$$w^{km} = \frac{p(m|C_k)\ P(C_k)}{P(m)} = p(C_k|m) \tag{16}$$

In RBF network, an unsupervised learning procedure is adopted for choosing the parameters for the radial basis functions. And the computation of output layer weights is formulated as a quadratic optimization problem [29].

4.5 BAYESIAN CLASSIFIER

In Bayesian classifier, the probability that one particular data point $\mathbf{y}_i$ belongs to class C_k is denoted as $P(C_k|\mathbf{y}_i)$ and is computed using Bayes' theorem [24, 29]

$$P(C_k|\mathbf{y}_i) = \frac{p(\mathbf{y}_i|C_k)P(C_k)}{p(\mathbf{y})} \tag{17}$$

where $P(C_k)$ is the class prior probability, which can be estimated from the training dataset, $p(\mathbf{y}_i|C_k)$ is the conditional probability density function describing the probability distribution of data point $\mathbf{y}_i$ inside class C_k, and $p(\mathbf{y}) = \sum_{k=1}^{T} p\{\mathbf{y}_i|C_k\}P(C_k)$ is the scaling factor. $p(\mathbf{y}_i|C_k)$ can be approximated using the Gaussian mixture model (GMM) as

$$p(\mathbf{y}_i|C_k) \approx \sum_{j=1}^{\xi} p(\mathbf{y}_i|j)\ P(j) \tag{18}$$

where ξ is the number of Gaussian mixture components, $p(\mathbf{y}_i|j)$ is the probability density function of the mixture component, which is a Gaussian distribution with mean $\boldsymbol{\mu}_j$ and covariance $\boldsymbol{\Sigma}_j$, and $P(j)$ is the weight of the mixture component and it satisfies $0 < P(j) < 1$ and $\sum_{j=1}^{\xi} P(j) = 1$. The GMM parameters, i.e. $\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_\mathbf{j}, \boldsymbol{\Sigma}_\mathbf{j}, P(\mathrm{j})\}$ can be calculated by using expectation maximization (EM) algorithm on the training dataset [24]. After obtaining the GMM parameters, the new data point can be assigned to the class, to which this data point has the largest probability.

4.6 FUZZY SUPPORT VECTOR MACHINE (FSVM)

The support vector machine (SVM) transforms the data points from the original input space into a higher dimensional space, which is defined by a kernel function. Then SVM searches for a hyperplane that maximizes the margin between different groups of data points in this space (see Figure 1d). FSVM is a variant of SVM and it is effective in dealing with the dataset contaminated by noises [26, 32].

In FSVM each data point is assigned a weight and these weights are augmented to the original training dataset $\mathbf{Y}$, which then becomes $\{\mathbf{y}_k, C_k, \rho_k\}$, where $0 < \rho_k < 1$ are the weights describing the degree of data point $\mathbf{y}_k$ belonging to class C_k. The sample weights ρ_k are computed based on an exponentially decaying function [32].

$$\rho_k = \frac{2}{1+exp(\beta a_k^{cen})} \quad \beta \in [0,1] \tag{19}$$

where $d_k^{cen} = \|\mathbf{y}_k - \mathbf{v_i}\|^{1/2}$ is the Euclidean distance between the data point $\mathbf{y}_k$ and its cluster centre $\mathbf{v_i}$, and β determines the steepness of the decay. The cluster centre is the geometrical mean of the data points belonging to the same class.

FSVM algorithm transforms input data to a higher dimensional space through a nonlinear function $\emptyset(\mathbf{y})$. Then it searches for an optimal separation hyperplane in that space by solving a quadratic programming (QP) problem [26]

$$\text{Minimize} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + G\sum_{k=1}^{N}\rho_k\,\xi_k \tag{20}$$

$$\text{Subject to } z_k(\mathbf{w}^T \cdot \emptyset(\mathbf{y_k}) + \boldsymbol{b}) \geq 1 - \xi_k \quad k = 1, \dots, N \tag{21}$$

$$\text{and} \quad \xi_k \geq 0 \quad k = 1, \dots, N \tag{22}$$

where the pair ($\mathbf{w}$, b) defines the separation hyperplane, in which $\mathbf{w}$ is a normal vector of the hyperplane and b is a bias. G is the regularization parameter to balance the margin maximization and misclassification, and $\xi_k \geq 0$ is the error term due to the misclassification. The above QP problem can be transformed into its dual form [26]

$$\text{Maximize } \sum_{k=1}^{N}\alpha_k - \frac{1}{2}\sum_{k=1}^{N}\sum_{j=1}^{N}\alpha_k\alpha_j\, z_k z_j K(\mathbf{y_k}, \mathbf{y_i}) \tag{23}$$

$$\text{Subject to } \sum_{k=1}^{N}\alpha_k z_k = 0 \tag{24}$$

$$\text{and} \quad 0 \leq \alpha_k \leq \rho_k G \tag{25}$$

where α_k is the Lagrange multiplier, and $K(\mathbf{y_k}, \mathbf{y_i})$is the kernel function in the form of $K(\mathbf{y_k}, \mathbf{y_i}) = \emptyset(\mathbf{y_k})^T\emptyset(\mathbf{y_i})$. The Gaussian kernel $K(\mathbf{y_k}, \mathbf{y_i}) = e^{-\gamma\|\mathbf{y_k}-\mathbf{y_i}\|^2}$ is adopted in this paper, where γ is the variance parameter. After obtaining the maximized hyperplane, FSVM can predict the class label z_* for a new data point $\mathbf{y}_*$ as

$$z_* = \text{sgn}[\textstyle\sum_{k=1}^{N}\alpha_k z_k K(\mathbf{y_k}, \mathbf{y}_*) + b] \tag{26}$$

5 EXPERIEMNTAL SETUP

The experiment setup for PD measurement is depicted in Figure 2. A commercially available equipment (Omicron MPD600), which complies with IEC60270 standard, was used in the experiment. Artificial PD models were designed to simulate five types of PD sources including corona, discharge in oil, surface discharge, internal discharge, and discharge due to floating particles (Figure 3). In the remaining of this paper, the terms of PD sources and PD classes are both referred to the PD models in Figure 3 and thus interchangeable.

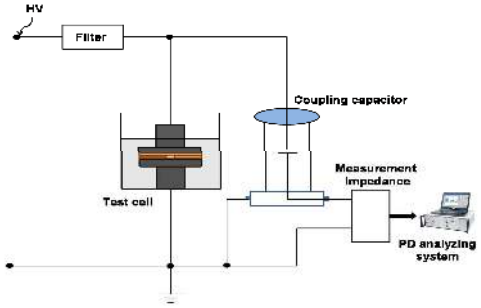


Figure 2. PD measurement setup.

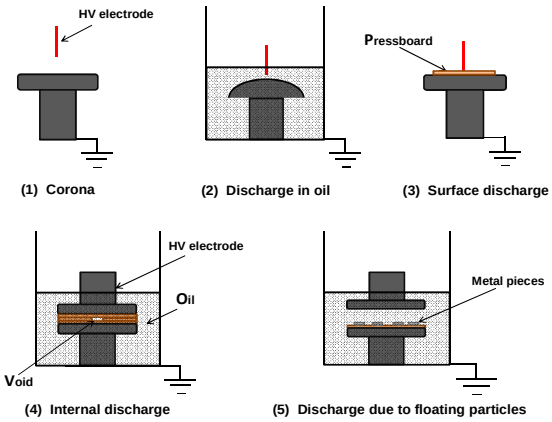


Figure 3. PD models used in the experiments.

For each of the five PD models in Figure 3, measurements were conducted under three different applied voltage levels (above the PD inception voltages) and three different noise levels. At one acquisition, PD pulses of 100 power cycles were recorded. For each of five PD models, 200 acquisitions were recorded and thus total 1000 data points were used to build up a database for evaluating the algorithms presented in this paper. Figure 4 shows a typical three dimensional (3D) PD patterns of the above five models.

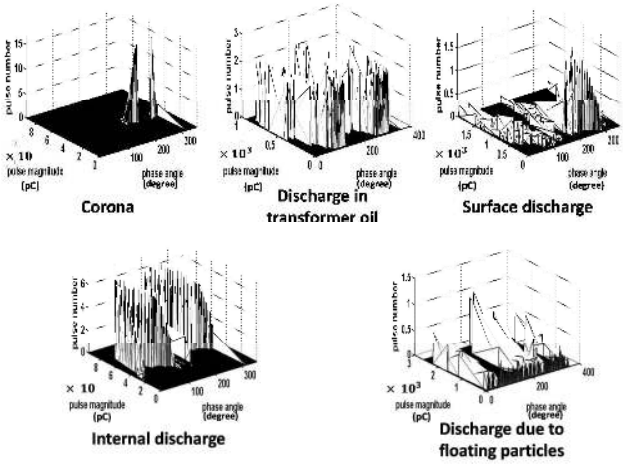


Figure 4. Typical PD patterns of five PD models.
(The test voltages for these models are in the range of 6 and 9 kV)

6 RESULTS AND ANALYSIS

This section presents PD source classification results and provides a comparison on the performance of various feature extraction and pattern recognition algorithms. In algorithmic implementations some software routines from the available toolbox are adopted with necessary modifications and extensions [28, 33].

6.1 DATA PREPROCESSING

As mentioned in Section 5, the original dataset obtained from PD measurement consists of 1000 data points. And each data point is made up of PD pulses of 100 power cycles. To perform PD source classification using the feature extraction

approaches and pattern recognition algorithms presented in Sections 3 and 4, the original dataset needs to be converted to the phase-revolved dataset (i.e. dataset **X**) consisting of average discharge pulse magnitude, maximum discharge pulse magnitude, and discharge number count in each of the 200 phase windows. Then the phase resolved dataset is fed into feature extraction algorithms to form a feature reduced dataset (i.e. dataset **Y**), which will then be used by pattern recognition algorithms for PD source classification. For the pattern recognition algorithms with DWT, the original dataset is adopted and then its feature size is reduced by using first four moment statistics (Section 3.5). The information regarding the feature size of different dataset is presented in Table 1.

Table 1. Information of feature size of datasets.

Feature extraction	Information of dataset **X**	Feature size of dataset **X**	Feature size of dataset **Y**
Statistic operator	Phase resolved PD data in 200 phase windows	600	24
PCA	Phase resolved PD data in 200 phase windows	600	7
KPCA	Phase resolved PD data in 200 phase windows	600	7
SNE	Phase resolved PD data in 200 phase windows	600	7
DWT	PD pulses of 100 power cycles	around 40,000	36

Note: dataset **X** is the phase resolved dataset while dataset **Y** is the feature reduced dataset.

To evaluate the pattern recognition algorithms, dataset **Y** is randomly split into two parts: a training dataset (consisting of about 70% data points of dataset **Y**) and a testing dataset (consisting of about 30% data points of dataset **Y**). The training dataset is used to find the model parameters in FSVM and ANNs. These parameters include the number of neighbors in *K*NN, the number of hidden nodes in MLP and RBF, the number of mixture components in Bayesian classifiers, G and γ (G is the regularization parameter and γ is the variance parameter of the Gaussian kernel) of FSVM. To decide the optimal values of the above parameters, the ten-fold cross validation on the training dataset is conducted. The process is as follows: the training dataset (consisting of l samples) are divided into 10 small datasets of size l /10. The algorithm is trained on nine datasets and tested on one dataset. It is repeated for 10 times and the mean accuracy is taken as the classification accuracy of the algorithm.

Once the optimal parameters are found, algorithms are trained on the full training dataset. Finally, the trained algorithms are applied to classify the data points in the testing dataset into one of PD sources (classes). The above dataset splitting, ten-fold cross validation, and testing are repeated 20 times for each algorithm, and the classification accuracy averaged over 20 trials will be recorded for comparisons.

6.2 SINGLE PD SOURCE CLASSIFICATION

Figure 5 presents the first three principle components of PCA on the phase resolved PD dataset **X** (Table 1). It can be seen that data points belonging to different PD classes scatter in the data space and mix up with each other. This implies the necessities of adopting algorithms with nonlinear discriminability for successful PD source classification. Table 2 to 6 summarizes the classification results of six pattern recognition algorithms integrated with different feature extraction approaches. Each table presents the overall classification accuracy rate (in percentage) and the classification accuracy rate (in percentage) with respect to each type of PD sources.

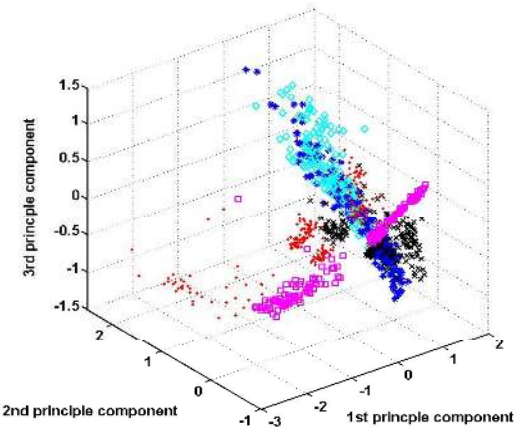


Figure 5. The first three principle components of phase resolved PD dataset ('.'- corona, '*'-discharge in oil, 'x'- surface discharge, □ -internal discharge, and ◇ - discharge due to floating particles)

From Tables 2 to 6, it can be seen that FSVM outperforms other five ANN algorithms. Amongst ANN algorithms, MLP also attains desirable and consistent classification accuracy. It is also not strange that the Two-layer network attains lower classification accuracy since its simple architecture may not be able to capture the nonlinearity of dataset. It is interesting to see that *K*NN performs well with the features extraction approaches of PCA, KPCA, and DWT. The similar phenomenon of *K*NN has also been reported in [24].

Tables 2 to 6 also reveals that *K*NN, Two-layer network, MLP, RBF, and FSVM integrated with DWT attain higher classification accuracy compared to these algorithms integrated with other four feature extraction approaches. *K*NN, MLP, RBF, Bayesian classifier, and FSVM integrated with KPCA and PCA attain comparable classification accuracy as they integrated with statistic operators. It is worthy to mention that KPCA and PCA only use 7 features, while statistic operators approach uses 24 features. The pattern recognition algorithms integrated with SNE could not attain desirable classification accuracy. The Two-layer network integrated with SNE even fails to make classifications. This might be due to the optimization in SNE is trapped into local minima in the cost function [27].

Table 2. Classification results of algorithms (with statistic operators).

Algorithm	Overall Rate	Class 1 Rate	Class 2 Rate	Class 3 Rate	Class 4 Rate	Class 5 Rate
*K*NN	90.8	99.4	83.8	86.7	99.5	85.2
Two-layer	90.6	94.7	87.2	86.1	97.6	88.1
RBF	92.5	98.5	87.4	90.1	97.8	89.2
MLP	92.0	98.1	84.8	91.6	99.1	86.6
Bayesian	93.2	97.5	85.6	91.0	99.3	93.3
FSVM	**95.1**	99.5	92.0	92.2	99.4	92.7

Note: Class 1 – corona; Class 2 – discharge in oil; Class 3 – surface discharge; Class 4 – internal discharge; Class 5 – discharge due to floating particle. Classification rates are in percentage.

Table 3. Classification results of algorithms (with DWT).

Algorithm	Overall Rate	Class 1 Rate	Class 2 Rate	Class 3 Rate	Class 4 Rate	Class 5 Rate
*K*NN	94.2	99.2	91.4	96.0	98.3	85.4
Two-layer	92.6	93.3	80.8	94.2	95.3	91.8
RBF	97.3	98.8	97.5	97.2	98.3	94.3
MLP	98.5	97.9	99.6	99.4	98.2	97.2
Bayesian	92.0	92.8	94.2	87.8	99.7	85.8
FSVM	**98.8**	97.4	99.4	98.9	100.0	98.3

Table 4. Classification results of algorithms (with PCA).

Algorithm	Overall Rate	Class 1 Rate	Class 2 Rate	Class 3 Rate	Class 4 Rate	Class 5 Rate
*K*NN	93.7	100.0	86.4	98.1	100.0	83.3
Two-layer	81.4	100.0	69.7	75.9	100.0	61.0
RBF	91.2	93.9	86.5	91.7	99.8	84.5
MLP	93.6	100.0	83.3	97.8	99.8	81.2
Bayesian	90.1	99.9	88.2	90.6	100.0	74.3
FSVM	**96.5**	99.9	93.4	99.7	99.8	89.1

Table 5. Classification results of algorithms (with KPCA).

Algorithm	Overall Rate	Class 1 Rate	Class 2 Rate	Class 3 Rate	Class 4 Rate	Class 5 Rate
*K*NN	93.9	100.0	90.3	96.0	99.8	82.8
Two-layer	83.2	98.9	74.2	82.8	99.3	59.6
RBF	91.7	96.9	87.8	91.7	100.0	82.0
MLP	93.8	100.0	87.9	99.0	99.5	81.3
Bayesian	92.4	99.9	87.9	95.3	99.1	78.4
FSVM	**96.1**	99.8	92.5	99.7	99.9	88.1

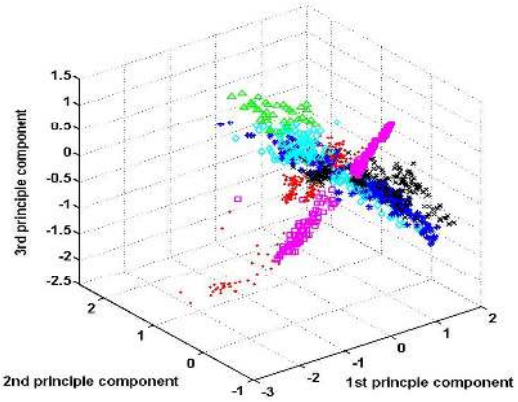
Table 6. Classification results of algorithms (with SNE).

Algorithm	Overall Rate	Class 1 Rate	Class 2 Rate	Class 3 Rate	Class 4 Rate	Class 5 Rate
*K*NN	87.2	97.6	79.4	85.8	93.2	79.9
Two-layer	28.7	43.2	55.1	6.9	34.7	0
RBF	85.2	92.4	76.7	82.5	91.7	83.7
MLP	89.4	99.4	85.7	83.3	99.3	79.3
Bayesian	85.9	94.1	83.8	80.0	94.5	77.0
FSVM	**92.3**	100.0	86.3	91.3	99.2	84.5

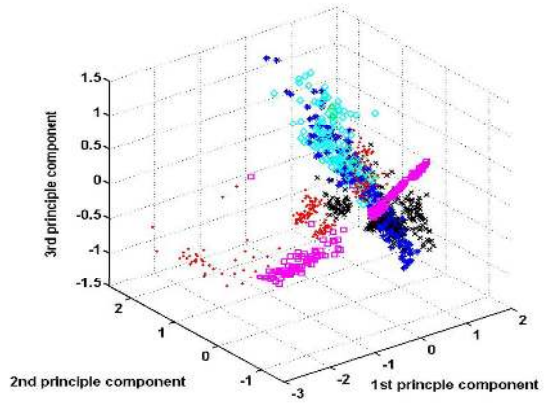
6.3 MULTIPLE PD SOURCES CLASSIFICATION

To investigate the approaches to multiple PD sources classification, two artificial datasets are generated. The first dataset is generated by firstly superposing the data points of PD model 4 (internal discharge) and 5 (discharge due to floating particles), and then combining these data points with dataset **X** to form a six-class dataset. The second dataset is generated by firstly superposing the data points of PD model 2 (discharge in oil) and 5, and then combining these data points with dataset **X** to form a six-class dataset. Figure 6 shows the first three components of PCA on the above two six-class datasets. The classification results on these two six-class datasets are summarized in Tables 7 and 8. All five PD models (Figures 3) were configured using a same test cell. By adjusting the geometric configurations, the PD inception voltages of different PD models are in the closed range. And the testing voltages for PD data acquisition were set slightly above the inception voltages. Therefore, the superposed PD data of different PD models were acquired under close range of testing voltages.

It can be seen from Table 7 that all six algorithms can successfully classify PD class 6(a), which is obtained by superposing the data points of PD model of internal discharge and those of PD model of discharge due to floating particles. In contrast, as shown in Table 8, the six algorithms attain lower classification accuracy on PD class 6(b), which is obtained by superposing the data points of PD model of discharge in oil and those of PD model of discharge due to floating particles. This can be explained by Figure 6. In Figure 6a the data points of PD class 6a (in green color triangle) are relatively separated from the data points of other five classes. However, in Figure 6b the data points of PD class 6b (in green color triangle) are mixed up with the data points of other classes and this causes difficulties in classification.



(a) Dataset X and the data points obtained from the superposition of PD models 4 and 5 (1. '.'- corona, 2, '*'- discharge in oil, 3, 'x'- surface discharge, 4. □ - internal discharge, 5. ◇ - discharge due to floating particles, and 6. △- combination of 4 and 5)



(b) Dataset X and the data points obtained from the superposition of PD models 2 and 5 (1. '.'- corona, 2, '*'- discharge in oil, 3, 'x'- surface discharge, 4. □ - internal discharge, 5. ◇ - discharge due to floating particles, and 6. △- combination of 2 and 5)

Figure 6. The first three principle components of two six-class datasets.

Table 7. Classification results of algorithms (with statistic operators) on six-class dataset I (Figure 6a).

Algorithm	Overall Rate	Class1 Rate	Class2 Rate	Class3 Rate	Class4 Rate	Class5 Rate	Class6 (a) Rate
KNN	92.2	99.0	85.7	89.7	99.2	86.5	100.0
Two-layer	90.3	93.8	86.0	87.6	95.5	87.5	100.0
RBF	93.2	98.2	87.9	92.1	95.7	91.0	100.0
MLP	94.7	98.5	92.4	92.2	98.7	91.3	97.5
Bayesian	89.1	95.6	79.0	86.5	99.3	84.0	99.2
FSVM	**95.6**	99.3	91.8	93.8	99.8	93.3	98.3

Note: Class 1– corona; Class 2 – discharge in oil; Class 3 – surface discharge; Class 4 – internal discharge; Class 5 – discharge due to floating particle; Class 6 (a) – superposition of Class 4 and Class 5. Classification rates are in percentage.

Table 8. Classification results of algorithms (with statistic operators) on six-class dataset II (Figure 6b).

Algorithm	Overall Rate	Class1 Rate	Class2 Rate	Class3 Rate	Class4 Rate	Class5 Rate	Class6 (b) Rate
KNN	89.9	99.7	82.9	88.2	99.3	85.7	57.5
Two-layer	87.7	94.9	85.6	85.7	96.8	86.7	28.3
RBF	90.8	99.0	81.9	91.0	96.8	91.0	63.3
MLP	92.5	98.1	87.5	92.5	98.5	94.0	50.8
Bayesian	86.7	95.3	75.1	87.1	99.2	81.5	65.8
FSVM	**94.7**	99.2	90.6	94.4	99.7	94.5	64.2

Note: Class 1 to Class 5 is the same as in Table 7; Class 6 (b) – superposition of Class 2 and Class 5. Classification rates are in percentage.

Another approach for recognizing multiple PD sources is the probability estimation. For the above data points obtained by superposition, their probabilities with respect to each single PD source are computed using FSVM algorithm (with statistic operators). Tables 9 and 10 show the calculated probabilities of ten data points (used as examples) in the above two multiple PD source datasets.

Table 9. Calculated class probability of multiple PD source (superposition of PD model 4 and model 5).

samples	Class 1 Probability	Class 2 Probability	Class 3 Probability	Class 4 Probability	Class 5 Probability
1	4.5	19.1	0.9	2.7	**72.8**
2	8.9	21.4	2.9	3.2	**63.6**
3	8.7	24.5	4.3	2.3	**60.2**
4	8.8	23.7	2.2	3.0	**62.3**
5	13.1	29.8	3.9	5.1	**48.1**
6	0.2	2.9	0	0.6	**96.3**
7	1.8	3.3	1.1	3.8	**90.0**
8	0.4	0 .4	0.8	0.4	**98.0**
9	1.0	2.8	0.7	1.5	**94.0**
10	14.6	31.3	4.2	5.7	**44.2**

Table 10. Calculated class probability of multiple PD source (superposition of PD model 2 and model 5).

samples	Class 1 Probability	Class 2 Probability	Class 3 Probability	Class 4 Probability	Class 5 Probability
1	11.4	28.5	6.4	1.8	**51.9**
2	0.3	**53.2**	15.7	0.3	30.5
3	0	**92.6**	7.1	0	0.3
4	12.1	**60.7**	2.3	0.9	24.0
5	20.1	26.6	3.6	0.5	**49.2**
6	0.9	14.3	0.7	0.7	**83.4**
7	0	**86.8**	6.2	0.2	6.8
8	0	**92.5**	0	7.1	0.4
9	0.6	27.8	22.7	0.5	**48.4**
10	1.7	47.5	0.6	0	**50.2**

It can be seen that from Table 9 that the largest probabilities are of PD class 5 (discharge due to floating particles). This coincides with the fact that the 10 data points are generated by superposing PD model 4 (internal discharge) and model 5. However, Table 9 also indicates that the second largest probabilities are of PD class 2 (discharge in oil) instead of PD class 4. This is contradictory to the fact that these data points are obtained by combining PD model 4 and 5. This might be due to some extents of similarities between the geometric configuration of PD model 5 and PD model 2 (refer to Figure 3). Such similarities may cause the ambiguity in classification.

Table 10 shows that the first and second largest probabilities are of PD Class 2 or 5. It coincides with the fact that these data points are generated by superposing PD model 2 and 5. This implies that the probability approach can recognize multiple PD source of PD model 2 and 5. However, due to the complexity of multiple PD source classification, further research is still required for investigating the appropriate approaches for identifying and separating multiple PD sources that may occur in HV equipment simultaneously.

6.4 STATISTICAL COMPARISON OF RESULTS

This section adopts the Friedman statistic test [30] to provide a quantitative performance comparison of different pattern recognition algorithms. It first ranks each algorithms in the above seven tables (Tables 2 to 8). Based on the overall classification accuracy of each table, the best-performing algorithm would be assigned with rank 1, the second best one would be assigned with rank 2, and so on. The average ranks of algorithms are computed as $R_j = \frac{1}{N}\sum_i r_i^j$, where r_i^j is the rank of the j-th algorithm on the i-th table. The averaged ranks of the pattern recognition algorithms developed in this paper are: *K*NN 3.4, Two-layer network 5.6, RBF 3.6, MLP 2.6, Bayesian 4.7, and FSVM 1.

To formally state that FSVM is the best among all six algorithms, it is necessary to reject the null-hypothesis, which states that, all the algorithms are equivalent so that any differences among their average ranks R_j are merely random. This can be done by computing the Friedman statistic [30]

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \tag{27}$$

$$\chi_F^2 = \frac{12N}{k(k+1)} \times \left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{28}$$

Based on the average ranks of six algorithms, the F_F is calculated as 13.92, which is greater than the critical value of $F(5,35)=2.49$ at significant level $\alpha = 0.5$. Therefore, the above null-hypothesis can be rejected. And this indicates that the performances of six algorithms are not equivalent at the significant level of $\alpha = 0.5$.

Then, Benferroni–Dunn test is used to compare the FSVM algorithm to the other algorithms. In this test, a critical difference (CD) value is defined as $CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}}$ where q_α values is 2.055 at $\alpha = 0.5$ [24]. If the rank difference between FSVM and another algorithm is at least this critical difference value, then it can be considered that FSVM is significantly superior to that algorithm. The differences between FSVM and *K*NN, two-layer network, RBF, and Bayesian classifier are greater than CD value. Thus it is able to claim that FSVM is significantly better than the other four algorithms. The difference in average ranks between FSVM and MLP is 1.6, which is less than the CD value of 2.055. However, the classification results in Tables 2 to 8 all indicate that FSVM outperforms MLP. The reasons of FSVM outperforms the ANNs would be the capabilities of the FSVM solving classification problem in higher dimensional space and dealing with the problem of noises and outliers of datasets.

7 DISCUSSION

This section discusses some issues of meaningful application of the proposed pattern recognition algorithms to practical PD source classification. The adaption and extension of the proposed algorithms for concurrent multiple PD sources classification will also be briefly discussed in this section.

7.1 PRACTICAL ISSUES OF THE APPLICATION OF PROPOSED ALGORITHMS

As mentioned in Section 4, by making use of a training database, the pattern recognition algorithms can learn the mathematical relationship between PD measurement data points and their corresponding types of PD sources. After the learning (training) phase, in the testing phase the algorithms are used to classify any PD measurement data in the testing dataset (not included in the training database) into one type of PD sources, which are defined in the training database.

PD phenomena are stochastic in nature. PD measurements are influenced by many factors such as ageing mechanisms of insulation system, PD signals distortion and attenuation, interference and noise, and multiple PD sources existence. Therefore it is still a non-trivial task to apply the proposed algorithms to practical PD sources classification of HV equipment. Two difficult scenarios are: (1) the PD sources of the testing dataset are not the exactly same as those in the original training database; and (2) the PD sources of the testing dataset are even not the part of the original training database [3], [10]. For these two scenarios, there is a certain degree of inaccuracy and uncertainty in determining the types of PD sources for PD data in the testing dataset.

Many algorithms including the conventional Bayesian classifier could not deal with the above two scenarios. Bayesian classifier has the limitation of assigning the pattern to the known classes when it is not part of the original patterns in the training database. As the consequence, it may make a false classification. This paper proposed a fuzzy support vector machine (FSVM, refer to Sections 4.6 and 6.3) algorithm, which can provide probabilistic interpretation when making classification on the unknown pattern. The following will discuss the application of FSVM to dealing with the uncertainties in the above two scenarios.

For the above first scenario, FSVM ccomputes the probabilities of PD measurement data (in the testing dataset) that they belong to each type of PD sources defined in the training database. Table 11 demonstrates such probability interpretation using FSVM algorithm proposed in this paper. In Table 11, the training and testing procedures are the same as those described in Section 6. The training database consists of 700 PD measurement data points of five PD sources, and the testing database consists of 300 data points. Due to the space limitation, Table 11 only presents the results of 15 data points of testing dataset.

It can be seen from Table 11 that each data point has different degree of probability belonging to different type of PD sources. For example, data point 5 has about 88% possibility of belonging to the PD source of discharge in oil and has about 11% possibility of belonging to the PD source of surface discharge. And data point 15 has about 23% probability of belonging to PD source of discharge in oil and has about 74% probability of belonging to PD source of discharge due to floating particles. Such probability interpretation might be beneficial to the practical PD sources classification using the proposed pattern recognition algorithms.

Table 11. FSVM classification results with probability for 15 PD measurement data points in the testing dataset (with statistic operators).

Data Point	Corona	Discharge in oil	Surface discharge	Internal discharge	Discharge due to floating particles
1	97%	1%	0	0	2%
2	91%	3%	2%	2%	2%
3	79%	5%	3%	8%	5%
4	0	93%	0	0	7%
5	0	88%	11%	1%	0
6	0	72%	27%	0	1%
7	1%	7%	91%	1%	0
8	0	32%	68%	0	0
9	0	4%	90%	5%	1%
10	2%	1%	5%	90%	2%
11	2%	1%	3%	94%	0
12	3%	0	4%	92%	1%
13	0	6%	5%	0	89%
14	3%	1%	0	0	96%
15	1%	23%	1%	1%	74%

Moreover, the probability in Table 11 can also provide a criterion to measure the similarity between the PD source in the testing dataset and that in the original training database. In this sense, the probabilistic interpretation may also be used for the second scenario, i.e. the types of PD sources of testing data are even not the part of the original training sources. In such scenario, the probability based interpretation can be used to measure the similarity between the PD sources of testing data and the already known PD sources in the original training database.

To achieve the desirable generalization capability of the proposed pattern recognition algorithms, the training database needs to include as many types of PD sources as possible. Readers may refer to Krivda's paper [3] for more information of constructing such training database. However, in practice the algorithms might be tasked to recognize the PD sources, which are not included in the training database. In this situation, the overall performance of the algorithms will be degraded. For example, two PD sources (corona and discharge in oil) are used to train FSVM and five PD sources (corona, discharge in oil, surface discharge, internal discharge, and discharge due to floating particles) are used in the testing phase. The classification results show that FSVM can attain 95% classification rate for the testing data points belonging to corona and discharge in oil; however, the data points belonging to other three PD sources are wrongly classified into PD sources of either corona or discharge in oil.

A two-stage procedure might be useful to handle the above situation, i.e. the algorithms trained to recognise A and B are used to recognize the neither pattern of A nor B (since the presented pattern was C or D or F). At the first stage, A and B is treated as a single type PD source and a one-class FSVM algorithm is adopted. This algorithm can detect whether the presented pattern is either A or B, or neither A nor B. If the presented pattern is either A or B, then at the second stage the

normal FSVM algorithm is used to decide the types of PD sources. If the presented pattern is neither A nor B, then it can tell that "I don't know the presented pattern (it is not in the training database)" [3]. And at the second stage the normal FSVM is still used to indicate the similarity (in probability sense) between the presented pattern and pattern A and B to assist further investigation, i.e. it tells that "the presented pattern is not in the training database, but it has 10% similarity with pattern A and has 15% similarity with pattern B".

It is also worthy to mention that some other factors are also needed to be considered when applying the proposed algorithms to PD sources classification. For example, the resolution of PD measurement system has an influence on the overall performance of the algorithms. For example, if the algorithms are trained with PD signals with full resolution and tested on the squashed (distorted) PD signals, then the overall performance of the algorithms will be significantly decreased.

7.2 THE EXTENSION OF THE PROPOSED ALGORITHMS FOR MULTIPLE PD SOURCES SEPERATION AND CLASSIFICATION

The approaches reported in the literature for multiple PD sources separation and classification include mixed Weibull model [7], equivalent time and bandwidth computation of the acquired PD signal [9], auto-correlation function [18], and density based spatial clustering [23] … etc. These approaches are based on the assumptions that the PD signals generated by the same PD source have similar shapes. Current research is focusing on enhancing the PD pattern separation effectiveness and subsequently improving multiple PD sources classification accuracy.

The authors of this paper are currently investigating the extension of the proposed pattern recognition algorithms for multiple PD sources separation and classification. A hybrid algorithm is under development, which consists of a blind source separation module for PD sources separation at the raw signal processing level and a fuzzy *c*-means clustering-fuzzy support vector machine algorithm for clustering and PD sources classification at the data and information processing level.

8 CONCLUSIONS

This paper provided a systematic study of applying pattern recognition techniques to PD source classification. The challenging issues of PD source classifications are investigated in this paper. To provide a comprehensive performance evaluation of various algorithms, extensive laboratory experiments on different PD source models are conducted. Future research will focus on the investigation of multiple PD source classification by conducting both laboratory experiments and onsite field measurements.

REFERENCES

[1] F.H.Kreuger, *Partial Discharge Detection in HV Equipment*, London Butterworth, UK, 1989.

[2] E. Gulski and A. Krivda, "Neural network as a tool for recognition of partial discharges", IEEE Trans. Electr. Insul., Vol. 28, No.6, pp.984-1001, 1993.

[3] A. Krivda, "Automated Recognition of Partial Discharges", IEEE Trans. Dielectr. Electr. Insul., Vol. 2, No.5, pp.796-821, 1995.

[4] E.Gulski, "Digital Analysis of Partial Discharge", IEEE Trans. Dielectr. Electr. Insul., Vol. 2, pp. 822-837, 1995.

[5] L. Satish and W.S. Zaengl, "Can Fractal Features be Used for Recognizing 3-D Partial Discharge Patterns", IEEE Trans. Dielectr. Electr. Insul., Vol. 2, No.3, pp. 352-359, 1995.

[6] A. Krivda and D. Birtwhistle, "Recognition of Multiple Partial Discharge Patterns", Eleventh International Symposium on High Voltage Engineering (Conf. Publ. No. 467), Vol.5, pp. 17-20, 1999.

[7] A. Contin, G.C. Montanari, and C. Ferraro, "PD Source Recognition by Weibull Processing of Pulse-Height Distributions", IEEE Trans. Dielectr. Electr. Insul., Vol. 7, No. 1, pp.48-58, February, 2000.

[8] IEC International Standard 60270, High-Voltage Test Techniques - Partial Discharge Measurements, Int'l. Electrotechnical Commission (IEC), 3rd edition, 2000.

[9] A. Contin, A. Cavallini, G.C. Montanari, and G. Pasini, "Digital Detection and Fuzzy Classification of Partial Discharge Signals", IEEE Trans. Dielectr. Electr. Insul., Vol. 9, No.3, pp. 335-348, 2002.

[10] R. Bartnikas, "Partial Discharges: Their Mechanism, Detection and Measurement", IEEE Trans. Dielectr. Electr. Insul., Vol. 9, No.5, pp. 763-808, 2002.

[11] M.M.A. Salama and R. Bartnikas, "Determination of Neural-Network Topology for Partial Discharge Pulse Pattern Recognition", IEEE Trans. Neural Network, Vol. 13, No.2, pp. 446-456, 2002.

[12] G.C. Stone, "Partial Discharge Diagnostics and Electrical Equipment Insulation Condition Assessment", IEEE Trans. Dielectr. Electr. Insul., Vol. 12, No.5, pp. 891-903, 2005.

[13] N.C. Sahoo, M.M.A. Salama, and R. Bartnikas, "Trends in Partial Discharge Pattern Classification: A Survey", IEEE Trans. Dielectr. Electr. Insul., Vol. 12, No.2, pp. 248-264, 2005.

[14] H. Zhang, T.R. Blackburn, B.T. Phung, and D. Sen, "A Novel Wavelet Transform Technique for On-line Partial Discharge Measurements, Part 1: WT De-noising Algorithm", IEEE Trans. Dielectr. Electr. Insul., Vol. 14, No. 1, pp. 3-14, 2007.

[15] A.M. Gaouda, A. El-Hag, T.K. Adbel-Galil, M.M.A. Salama, and R. Bartnikas, "On-Line Detection and Measurement of Partial Discharge Signals in a Noisy Environment", IEEE Trans. Dielectr. Electr. Insul., Vol. 15, No.4, pp. 1162-1173, 2008.

[16] S.M. Strachan, S. Rudd, S.D.J. McArthur, and M.D. Judd, "Knowledge-Based Diagnosis of Partial Discharges in Power Transformers", IEEE Trans. Dielectr. Electr. Insul., Vol. 15, No.1, pp. 259-268, 2008.

[17] A. Rizzi, F. Mascioli, F. Baldini, C. Mazzetti, and R. Bartnikas, "Genetic Optimization of a PD Diagnostic System for Cable Accessories", IEEE Trans. Power Delivery, Vol. 24, No.3, pp. 1728-1738, 2009.

[18] A. Contin and S. Pastore, "Classification and Separation of Partial Discharge Signals by Means of Their Auto-correlation Function Evaluation", IEEE Trans. Dielectr. Electr. Insul., Vol. 16, No.6, pp. 1609-1622, 2009.

[19] H. Al-Marzouqi and A. Contin, "Separation of Multiple Sources in PD Measurements Using an Intensity Based Clustering Algorithm", 2010 IEEE International Symposium on Electrical Insulation, pp.1-4, San Diego, USA 2010.

[20] N.Kuljaca, S.Meregalli, A.Contin, and A.Ukovich, "Separation of Multiple Sources in PD Measurements Using An Amplitude-Frequency Relation Diagram", IEEE 10th Int'l. Conf. Solid Dielectrics (ICSD), pp.1-4, Potsdam, Germany, 2010.

[21] D. Evagorou, A. Kyprianou, P.L. Lewin, A. Stavrou, V. Efthymiou, A. C.Metaxas, and G.E.Georghiou, "Feature Extraction of Partial Discharge Signals Using the Wavelet Packet Transform and Classification with a Probabilistic Neural Network", IET Sci. Meas. Technol., Vol.4, pp. 177-192, 2010.

[22] L. Hao and P.L. Lewin, "Partial Discharge Source Discrimination Using a Support Vector Machine", IEEE Trans. Dielectr. Electr. Insul., Vol. 17, No.1, pp. 189- 197, 2010.

[23] L. Hao, P.L. Lewin, J.A. Hunter, D.J. Swaffield, A. Contin, C. Walton, and M. Michel, "Discrimination of Multiple PD Sources Using Wavelet Decomposition and Principal Component Analysis", IEEE Trans. Dielectr. Electr. Insul., Vol. 18, No.5, pp. 1702-1711, 2011.

[24] C.M.Bishop, *Neural Networks for Pattern Recognition,* Oxford University Press, 1995.

[25] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A review", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, pp. 4-37, 2000.
[26] C.-F. Lin and S.-D. Wang, "Fuzzy Support Vector Machines", IEEE Trans. Neural Networks, Vol.13, pp. 464-471, 2002.
[27] G. Hinton and S. Roweis, "Stochastic Neighbor Embedding", Advances in Neural Information Processing Systems, Vol. 15, pp. 833-840, 2002.
[28] I.T.Nabney, *NETLAB: Algorithms for Pattern Recognition*, Springer, 2004.
[29] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
[30] J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets", J. Machine Learning Research, Vol.7, pp.1-30, 2006.
[31] D. Salomon, G. Motta, and D. Bryant, *Data compression: the Complete Reference* (4th edition), Springer, 2007.
[32] R. Batuwita and V. Palade, "FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalanced Learning", IEEE Trans. Fuzzy Systems, vol. 18, pp. 558-571, 2010.
[33] C.-C.Chang and C.-J.Lin, "LIBSVM: a library for support vector machines", available: http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2012.

ACKNOWLEDGMENT

The authors gratefully acknowledge Australian Research Council, Powerlink Queensland, Energex, Ergon Energy, and TransGrid for providing supports for this work. The authors also would like to thank the anonymous reviewers for their valuable comments and suggestions for improving the quality of this work.

Hui Ma (M'95) received his B.Eng. and M.Eng. degrees from Xi'an Jiaotong University, China in 1991 and 1994, M.Eng. (by research) degree from Nanyang Technological University, Singapore in 1998, and the Ph.D. degree from the University of Adelaide, Adelaide, Australia in 2008. Currently Dr. Ma is a research fellow in the School of Information Technology and Electrical Engineering, the University of Queensland, Australia. Prior to joining the University of Queensland, Dr. Ma has many years research and development experience. From 1994 to 1995, he was a researcher in Xi'an Jiaotong University, China. From 1997 to 1999, he worked as a firmware development engineer in CET Technologies Pte. Ltd., Singapore. He was with Singapore Institute of Manufacturing Technology as a research engineer from 1999 to 2003. His research interests include industrial informatics, condition monitoring and diagnosis, power systems, wireless sensor networks, and sensor signal processing.

Jeffery C. Chan (S'12) is currently pursuing a Ph.D. degree in the School of Information Technology and Electrical Engineering, the University of Queensland, Australia. He obtained his B.Eng. (Hons.) degree in Mechatronic Engineering and M.Phil. degree in Manufacturing Engineering and Engineering Management from City University of Hong Kong in 2006 and 2009 respectively. Before joining the University of Queensland, he worked for a number of projects on machine fault diagnosis and energy conservation for 3 years. His research interests include signal processing, partial discharge, and condition monitoring of power systems.

Tapan Kumar Saha (M'93-SM'97) was born in Bangladesh in 1959 and immigrated to Australia in 1989. He received his B.Sc Engineering (electrical and electronic) in 1982 from the Bangladesh University of Engineering & Technology, Dhaka, Bangladesh, M. Tech (electrical engineering) in1985 from the Indian Institute of Technology, New Delhi, India and the Ph.D. degree in 1994 from the University of Queensland, Brisbane, Australia. Saha is currently a Professor of Electrical Engineering in the School of Information Technology and Electrical Engineering, University of Queensland, Australia. Previously he has had visiting appointments for a semester at both the Royal Institute of Technology (KTH), Stockholm, Sweden and at the University of Newcastle (Australia). He is a Fellow of the Institution of Engineers, Australia. His research interests include condition monitoring of electrical plants, power systems and power quality.

Chandima Ekanayake (M'00) received his B.Sc.Eng.(Hons) in 1999 from University of Peradeniya Sri Lanka. He obtained his Tech. Lic. and PhD from Chalmers University of Technology Sweden in 2003 and 2006 respectively. Currently he is a lecturer in the School of Information Technology and Electrical Engineering, the University of Queensland (UQ), Brisbane, Australia. Before joining UQ he was with University of Peradeniya Sri Lanka as a Senior lecturer. During his PhD studies he was working for a European Union Project called REDIATOOL where he engaged in research related to Diagnostics of Transformer Insulation from dielectric response measurements. From 2001, he has been involving on condition monitoring of transformers installed at Ceylon Electricity Board, Sri Lanka. He was the Chair of IEEE Sri Lanka Section in year 2006 and 2007. His research interests are condition monitoring of power apparatus, alternatives for insulating oil, transient studies on power systems and energy related studies.